MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

AIM 416 May 1977

REPRESENTATION AND RECOGNITION OF THE SPATIAL ORGANIZATION OF THREE DIMENSIONAL SHAPES

D. Marr and H. K. Nishihara

Summary. The human visual process can be studied by examining the computational problems associated with deriving useful information from retinal images. In this paper, we apply this approach to the problem of representing three-dimensional shapes for the purpose of recognition.

1. Three criteria, *accessibility*, *scope & uniqueness*, and *stability & sensitivity*, are presented for judging the usefulness of a representation for shape recognition.

2. Three aspects of a representation's design are considered, (i) the representation's coordinate system, (ii) its primitives, which are the primary units of shape information used in the representation, and (iii) the organization the representation imposes on the information in its descriptions.

3. In terms of these design issues and the criteria presented, a shape representation for recognition should: (i) use an object-centred coordinate system, (ii) include volumetric primitives of varied sizes, and (iii) have a modular organization. A representation based on a shape's natural axes (for example the axes identified by a stick figure) follows directly from these choices.

4. The basic process for deriving a shape description in this representation must involve: (i) a means for identifying the natural axes of a shape in its image and (ii) a mechanism for transforming viewer-centred axis specifications to specifications in an object-centred coordinate system.

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-75-C-0643.

5. Shape recognition involves: (i) a collection of stored shape descriptions, and (ii) various indexes into the collection that allow a newly derived description to be associated with an appropriate stored description. The most important of these indexes allows shape recognition to proceed conservatively from the general to the specific based on the specificity of the information available from the image.
6. New constraints supplied by a conservative recognition process can be used to extract more information from the image. A relaxation process for carrying out this constraint analysis is described.

Introduction

Vision is a *process* that produces from images of the external world a description that is useful to the viewer and not cluttered by irrelevant information. One approach to understanding how this process works is through studying the information processing problems with which vision must deal. An important aspect of vision is concerned with the representation of three-dimensional shape. In this article, we shall argue that the way such information is represented is constrained by its applications and by the computational problems associated with deriving it from retinal images. These constraints are clarified and a representation consistent with them is presented.

Terminology

We shall reserve the term *shape* for the geometry of an object's physical surface. Thus two statues of a horse, cast from the same mould, have the same shape. A *representation* for shape is a formal scheme for describing shape or some aspects of shape, together with rules that specify how the scheme is applied to any particular shape. We shall call the result of using a representation to describe a given shape a *description* of the shape in that representation. A description may specify a shape only roughly, or in fine detail.

Issues raised by the representation of shape

There are many kinds of visually derivable information that play important roles in recognition and discrimination tasks. Shape information has a special character, because unlike colour or visual texture information, the representation of most kinds of information about shape requires some sort of coordinate system within which spatial relations can be described. For example, the information that distinguishes the different animal shapes in figure 1 is the spatial arrangement, orientation, and sizes of the sticks. Similarly, since left and right hands are reflexions of each other in space, any description of the shape of a hand that is sufficient for determining whether it is left or right must in some manner specify the relative locations of the fingers and thumb.

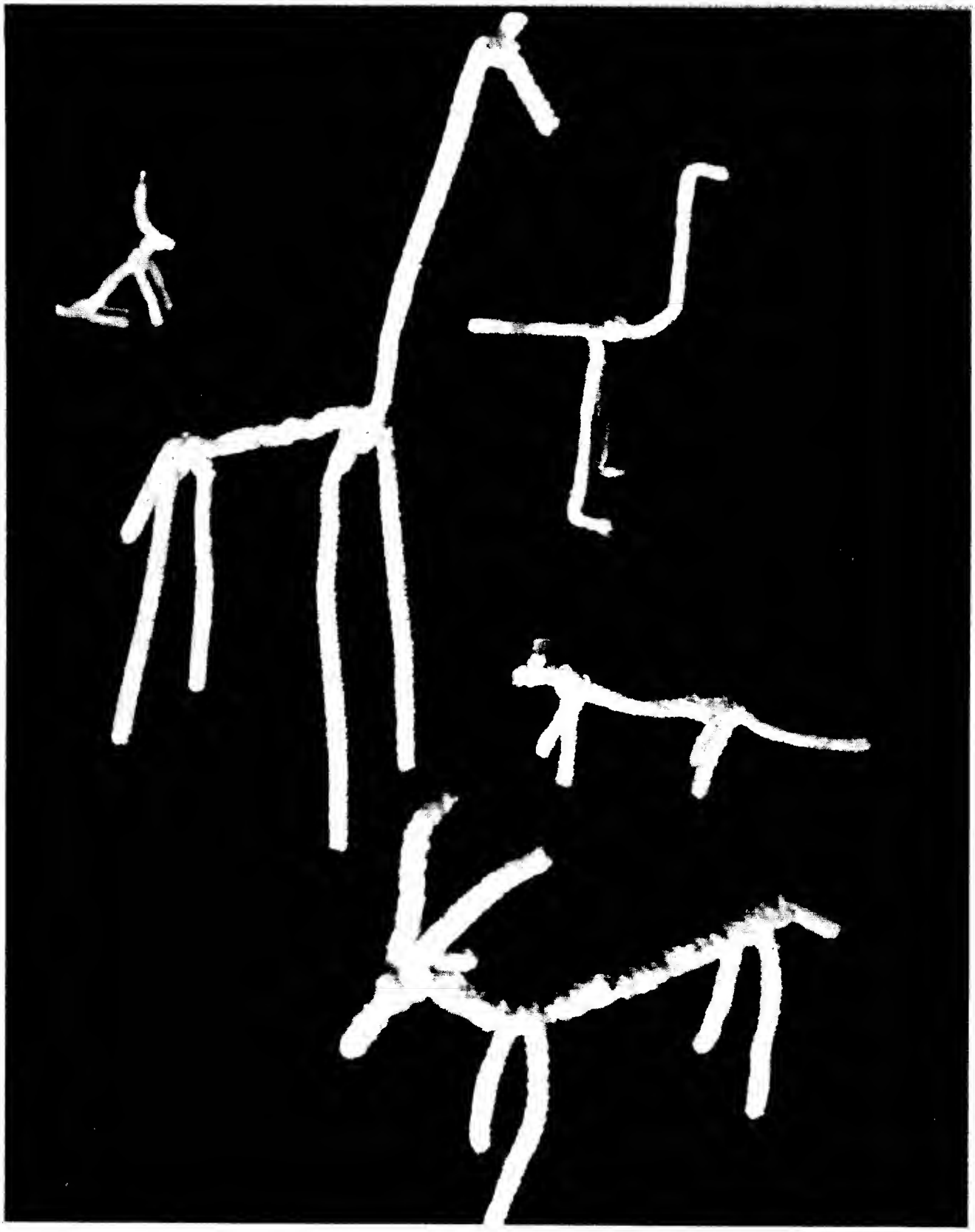
Criteria for judging the effectiveness of a shape representation

There are many different aspects of an object's shape, some more useful than others, and any one aspect can be described in a number of ways, for example by using different coordinate systems. Although it is difficult to formulate a completely general classification of shape representations, we attempt to set out here the main criteria by which they may be judged, and the basic design choices that have to be made when a representation is formulated.

(C1) *Accessibility*

Can the desired description be computed from an image, and can it be done reasonably inexpensively? There are fundamental limitations inherent in the information available in an image -- for example its resolution -- and the requirements of a representation have to fall within the limits of what is possible. Moreover, a description

Figure 1. These pipecleaner figures illustrate several of the points developed in this paper. A shape representation does not have to reproduce a shape's surface in order to describe it adequately for recognition; as we see here, animal shapes can be portrayed quite effectively by the arrangement and relative sizes of a small number of sticks. The simplicity of these descriptions is due to the correspondence between the sticks shown here and natural or canonical axes of the shapes described. To be useful for recognition, a shape representation must be based on characteristics that are uniquely defined by the shape and which can be derived reliably from images of it.



that is in principle derivable from an image may still be undesirable if its derivation involves unacceptably large amounts of memory or computation time.

(C2) *Scope and uniqueness*

What class of shapes is the representation designed for and do the shapes in that class have canonical descriptions in the representation? For example, a shape representation designed to describe planar surfaces and junctions between perpendicular planes would have cubical solids within its scope, but would be inappropriate for describing shapes with curved surfaces or needle-like protuberances. If the representation is to be used for recognition, it is also important that the description of a shape be unique, otherwise at some point in the recognition process, the difficult problem of deciding whether two descriptions describe the same shape would have to be addressed. If for example, one chose to represent shape using polynomials of degree n, the formal description of a given surface would depend on the particular coordinate system chosen. Since one would be unlikely to use the same coordinate system on two different occasions unless some additional conventions were being adhered to, even the same image of a surface could give rise to very different descriptions.

(C3) *Stability and sensitivity*

Within the above scope and uniqueness conditions, there lie questions about the continuity and resolution of a representation. To be useful for recognition, the degree of similarity between two shapes must be reflected in their descriptions, but at the same time even subtle differences must be expressible. These opposing conditions can be satisfied only if it is possible to decouple stable information, that captures the more general and less varying properties of a shape, from information that is sensitive to the finer distinctions between shapes. For example, consider a stick figure representation that uses the three-dimensional arrangement and size of stick elements to describe animal shapes, as in figure 1. The size of the sticks used gives one control over the stability and sensitivity of the resulting stick figure description. Stability is increased by using larger sticks; a single stick provides the most stable description of the whole shape, describing only its size and orientation. A description built of smaller sticks, on the other hand, would be sensitive to smaller, more local details such as the slight bends in an animal's limbs. Such details tend to be less stable, but can nevertheless be important for making fine distinctions between similar shapes.

Choices in the design of a shape representation

We now relate the effects of different choices in the design of a shape representation to the performance criteria listed above. Perhaps the most fundamental property of a representation is that it can make some types of information *explicit*, bringing the essential information to the foreground, and allowing the descriptions to be smaller and manipulated more quickly. Three aspects of a representation's design are considered here, (i) the representation's coordinate system, (ii) its primitives, which are the primary units of shape information used in the representation, and (iii) the organization the representation

imposes on the information in its descriptions.

(DI) *Coordinate systems*

The most important aspect of the coordinate system used by a representation is the way it is defined. If locations are specified relative to the viewer, we say the representation uses a *viewer-centred* coordinate system. If locations are specified in a coordinate system defined by the viewed object, the representation uses an *object-centred* coordinate system.

Viewer-centred descriptions are easier to produce but harder to use for recognition tasks than object-centred ones, because viewer-centred descriptions depend upon the vantage point from which they are built. As a result, any theory for recognition that is based on a viewer-centred representation must treat distinct views of an object essentially as distinct objects. The important characteristic of this approach is that it requires a potentially large store of descriptions in memory in exchange for a reduction in the magnitude and complexity of the computations that would otherwise be required to compensate for the effects of perspective. Minsky (1975) has suggested that this number might be minimized by an appropriate choice of shape primitives and of the views to be stored in memory. It is clear that much can be accomplished with this approach in some circumstances. For example, if squirrels need to distinguish trees from other objects but do not need to identify particular trees by their shape, they may be able to take advantage of some of the general characteristics of a vertical tree trunk's appearance that do not depend on the vantage point so long as it is on the ground nearby. In a representation based on these characteristics, all trees in the squirrel's environment would produce essentially the same description. For more complex recognition tasks, however, where finer subtleties of the arrangement of the object's components are important, it is unlikely that a viewer-centred representation can be found that will not be sensitive to the object's orientation. For example, consider the many orientation-dependent appearances of a human hand, that exist even if the fingers and thumb remain fixed with respect to each other. In order to distinguish a left hand from a right using a viewer-centred representation, one would probably have to treat this problem as many separate cases, one for each possible appearance of a hand.

The alternative to relying on an exhaustive enumeration of all possible appearances, is to use an object-centred coordinate system, placing greater emphasis on the computation of a canonical description which is independent of the vantage point. Ideally, only a single description of each object's spatial structure would then have to be stored in memory in order for that object to be recognizable from even unfamiliar vantage points. An object-centred description is however more difficult to derive since a unique coordinate system has to be defined for each object and that coordinate system has to be identified from the image before the description is constructed.

(D2) *Primitives*

The primitives of a representation are the most elementary units of shape information available in a representation, which is the type of information that the representation receives from earlier visual processes. For example, figure 2 illustrates an example of a representation whose primitives carry information about local surface orientation and distance (relative to the viewer) at thousands of evenly spaced locations in the visual field. We separate two aspects of a representation's primitives, the type of shape information they carry, which is important for questions of accessibility, and their size, which is important for questions of stability and sensitivity.

There are two principal classes of shape primitive, surface-based (two-dimensional) and volumetric (three-dimensional). Surface information is more immediately derivable from images. The simplest primitives useful for surface descriptions would specify just the location and size of small pieces of surface. More elaborate surface primitives like those used in figure 2 could include orientation and depth information as well. On the other hand, volumetric primitives carry information about the spatial distribution of a shape. This type of information is more directly related to the requirements of shape recognition than information about a shape's surface structure, and this often means that much shorter and therefore more stable descriptions can still satisfy the sensitivity criterion. The simplest volumetric primitive specifies just a location and an extent, and corresponds to a roughly spherical region in space. By adding a vector to this information, a roughly cylindrical region can be specified, where the length of the vector indicates the length of the cylinder and the spatial extent parameter indicates its diameter. A second vector could, in addition, indicate a rotational orientation about the first vector. This would make it possible to specify a pillow-shaped region whose cross section along the first vector is thicker in the direction of the second vector. An additional vector could also be used to specify a curvature in the axis of the cylindrical region by indicating its direction and magnitude.

The complexity of the primitives used by a representation is limited largely by the type of information that can be derived reliably by processes prior to the representation. While the information carrying capacity of primitives can be extended arbitrarily, there is a limit to the amount that is useful, since very detailed primitives will be derived less consistently by those earlier processes. In the extreme case, descriptions in a shape representation would consist of a single primitive. Such a representation would satisfy the uniqueness and stability conditions only if the information carried by the primitives were derived consistently by the processes supplying it. If this were so, however, those processes would already have accomplished shape recognition in specifying the primitive; and there would be no need for the representation.

Size is the other aspect of a representation's primitives that influences the information it makes explicit. In particular, information about features much larger than the primitives used is difficult to access since it is represented only implicitly in the configuration of a large number of smaller items. For example, consider how the arm of a human shape would be described in a surface representation like the one illustrated in figure 2. Only information about small patches of surface is present, so a rather sophisticated analysis of a large assembly of these is required to make explicit the presence

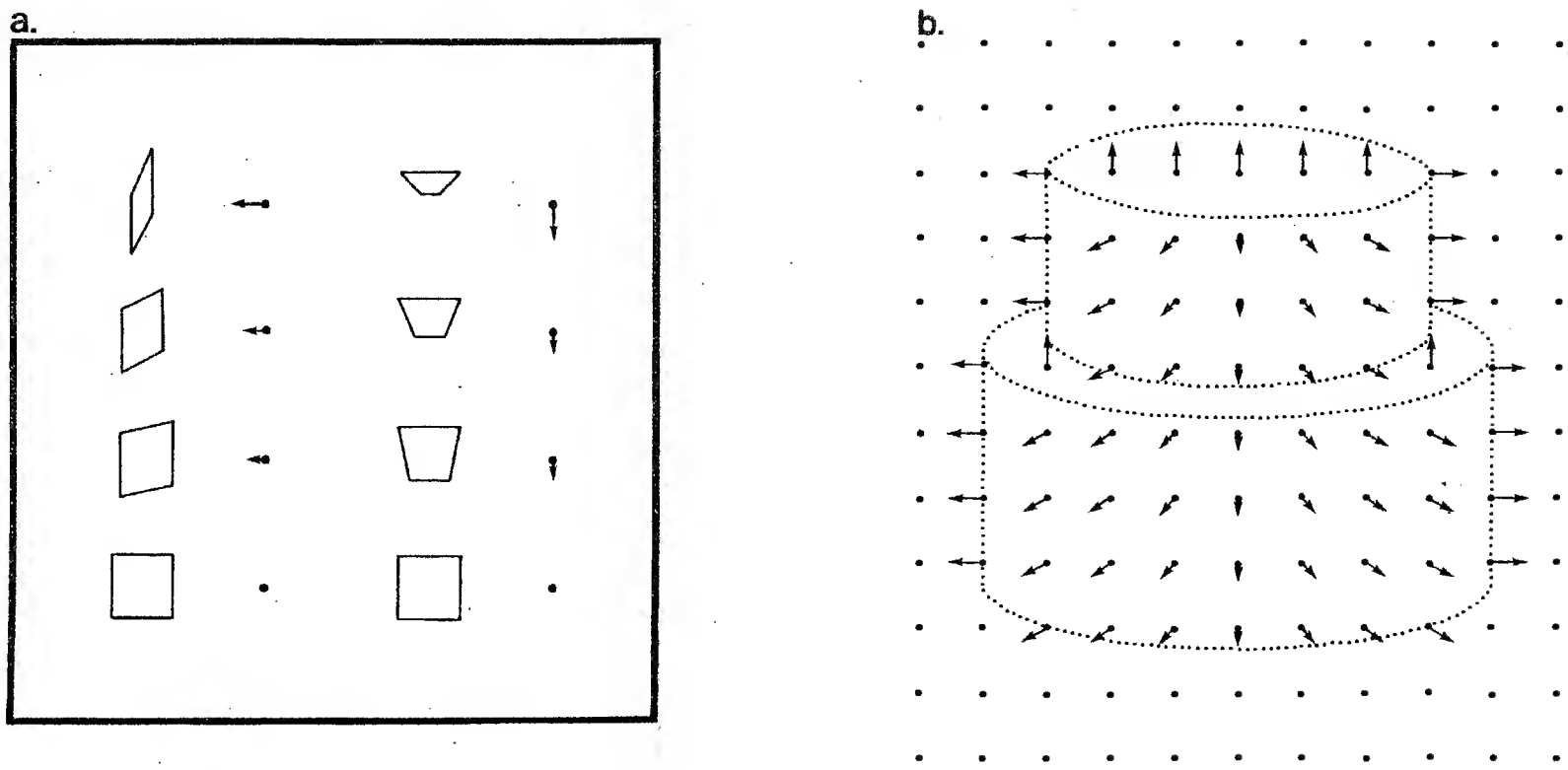


Figure 2. A description of an object's shape has to be derived via a description of its visible surface, since the information encoded in images, for example by stereopsis, shading, texture gradients, or occluding contours, is due to a shape's local surface properties. The objective of many early visual computations is to extract this information and we shall refer to the representation that makes it explicit as the $2\frac{1}{2}$-*D sketch*. An example is shown here which provides information about the orientation (relative to the viewer) of small patches of surface spaced evenly over the visual field. The arrows depict the direction of steepest decent and its magnitude by their direction and length (the longer the arrow, the greater the dip of the surface element out of the image plane). The information carried by this representation's primitives is therefore specified by these arrows.

of the arm-shape itself. A stick figure representation, on the other hand, can specify an arm explicitly with a single stick primitive of the appropriate size.

At the other end of the scale, features of a shape that are much smaller than the primitives used to describe it are not just inaccessible, they are completely omitted from the description. For example, the fingers of a human shape are not expressible in a stick figure description that uses only primitives the size of the arms and legs. Similarly, surface details much smaller than the primitives used in figure 2 would be inexpressible in that representation. Thus the size of the primitives used in a description acts as a very strong determiner of the kind of information made explicit by a representation, the information made available but not directly obtainable, and the information that is discarded.

(D3) *Organization*

The third design dimension we consider is the way shape information is organized by a representation. In the simplest case, no organization is imposed by the representation and all elements in a description have the same status. The local surface representation in figure 2 is an example. Alternatively, the primitive elements of a description can be organized into modules consisting, for example, of spatially adjacent elements of roughly the same size, in order to distinguish certain subgroupings of the primitives from others. This is closely related to the *principle of explicit naming* (Marr 1976) which states that it is important to be able to give names to groups of elements in a representation so that properties can be associated with them and external processes can reference them efficiently. Similar ideas occur in the fields of computer science and artificial intelligence (eg Winston 1975). A modular organization is especially useful for recognition because it can make sensitivity and stability distinctions explicit, by arranging for all constituents of a given module to lie at roughly the same level of stability and sensitivity.

The 3-D Model representation

In terms of the requirements of shape recognition, which we have attempted to quantify as the criteria C1, C2, and C3, a shape representation should be object-centred (D1), volumetric (D2), and modular (D3). These choices have strong implications, and a limited representation which we shall call the 3-*D model representation* can be defined quite directly from them.

Shapes having natural coordinate systems

Our first objective is to define a shape's object-centred coordinate system. If it is to be canonical it must be based on axes determined by salient geometrical characteristics of the shape, and conversely, the scope of the representation must be limited to those shapes for which this can be done. A shape's natural axes may be defined by elongation, symmetry or even motion (eg the axis of rotation), so that the coordinate system for a sausage should be defined by its major axis and the direction of its curvature, and that of a face by its axis of symmetry. Objects with many or poorly defined axes, like a sphere, a door, or a crumpled newspaper, will inevitably lead to ambiguities. For a shape as regular as a sphere this poses no great problem, because its description in all reasonable systems is

the same. For a door there are four distinguished axes, defined by the directions of its length, its width, its thickness, and also by the axis on which it is hinged. Since the number is small and doors are important, one could deal with each of the four possible descriptions of a door as separate cases. This would not be true for a crumpled newspaper, however, which is likely to have a large number of poorly defined axes.

At present, the problems we understand best are those surrounding the determination of axes based on a shape's elongation; so for this paper we shall limit the scope of the 3-D model representation to shapes whose natural axes are of this type. A large class of shapes that satisfy this condition are the *generalized cones*. (A generalized cone is the surface swept out by moving a cross-section of constant shape but smoothly varying size along an axis). Binford (1971) drew attention to this class of surfaces, suggesting that it might provide a convenient way of describing three-dimensional surfaces for the purposes of computer vision. We regard it as an important class not because their surfaces are conveniently described, but because such shapes have well-defined axes. Many common shapes are included in the scope of such a representation, because objects whose shape was achieved by growth are often described quite naturally in terms of one or more generalized cones. The animal shapes in figure 1 provide some examples -- the individual sticks are simply axes of generalized cones that approximate the shapes of parts of these animals.

Stick figure descriptions

To be useful for recognition the representation's primitives must also be associated with stable geometric characteristics. The natural axes of a shape satisfy this requirement, and we shall therefore base the 3-D model representation's primitives on them. A description that uses axis-based primitives can be thought of as a stick figure, like those depicted in figure 1. While only a limited amount of information about a shape is captured by such a description, that information is especially useful for recognition. We shall further limit the information carried by these primitives to just size and orientation information. This will enable us to develop the character of the 3-D model representation with a minimal commitment to inessential details. More elaborate details, such as curved axes or the tapering of a shape along the length of its axis, will not be included here.

The concept of a stick figure representation for shape is not new. Blum (1973) for example has studied a classification scheme for two dimensional silhouettes based on a *grassfire* technique for deriving a kind of stick figure from those shapes, and Binford (1971) introduced the generalized cone for three-dimensional shapes. These representations have one characteristic in common, however; they do not impose a modular organization on the information they carry. For example, each part of the arm of a human shape can correspond to at most one stick in these representations; it would not be possible to have both a single stick corresponding to the whole arm and three smaller sticks corresponding to the major segments of the arm in the same description.

Modular organization of the 3-D model representation

The modular decomposition of a description used for recognition must be well defined. This is best achieved, in the 3-D model representation as it is specified so far, by basing it on the canonical axes of a shape. Each of these axes can be associated with a coarse spatial context that provides a natural grouping of the axes of the major shape components contained within that scope. We shall refer to a module defined this way as a *3-D model*. Thus each 3-D model specifies:

(i) a *model axis*, which is the single axis that defines the extent of the shape context of the model. This is a primitive of the representation and it provides coarse information, such as size and orientation, about the overall shape described.

and optionally:

(ii) the relative spatial arrangement and sizes of the major *component axes* contained within the spatial context specified by (i). The number of component axes should be small and they should be of roughly the same size.

(iii) the names (internal references) of 3-D models for the shape components associated with the component axes, whenever such models have been constructed. Their model axes correspond to the component axes of this 3-D model.

Each of the boxes in figure 3 depicts a 3-D model with the model axis on the left and an arrangement of component axes on the right. The model axis of the human 3-D model makes explicit the gross properties (size and orientation) of the whole shape with a single primitive. The six component axes corresponding to the torso, head and limbs can each be associated with a 3-D model which would contain additional information about the decomposition of that component into an arrangement of smaller components. Although a single 3-D model is a simple structure, the combination of several in this kind of organizational hierarchy allows one to build up a description that captures the geometry of a shape to an arbitrary level of detail. We shall call such a hierarchy of 3-D models a *3-D model description* of a shape.

The example in figure 3 illustrates the important advantages of a modular organization for a shape description. The stability of the representation is greatly enhanced by including both large and small primitive descriptions of the shape and by decoupling local spatial relationships from more global ones. Without this modularization, the importance of the relative spatial arrangement of two adjacent fingers would be indistinguishable from that of the relationship between a finger and the nose. Modularity also allows the representation to be used more flexibly in response to the needs of the moment. For example, it is easy to construct a 3-D model description of just the arm of a human shape which could later be included in a new 3-D model description of the whole human shape. Conversely a rough but usable description of the human shape need not include an elaborate arm description. Finally, this form of modular organization allows one to trade-off scope against detail. This simplifies the computational processes that derive and use the representation, because even though a complete 3-D model description may be

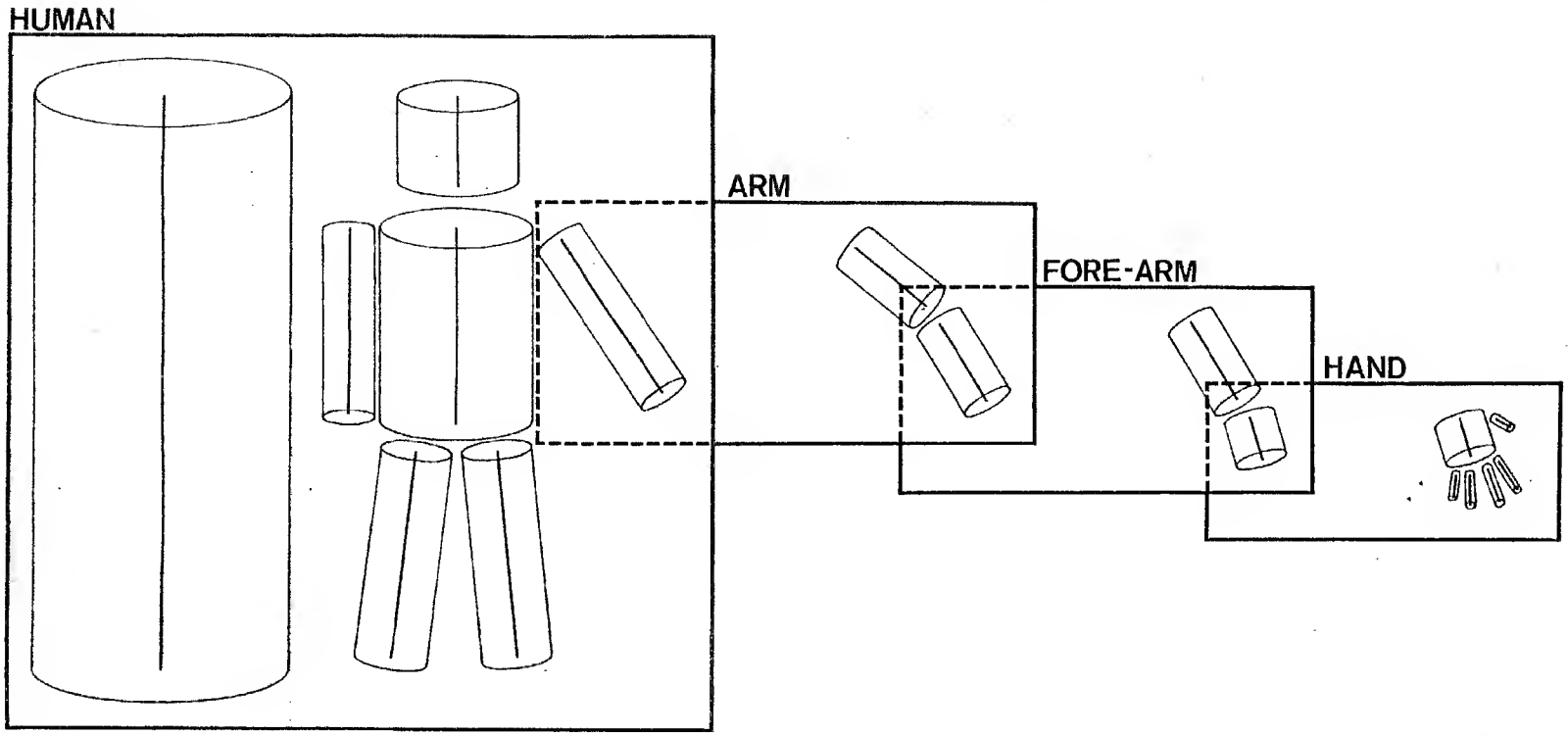


Figure 3. This diagram illustrates the organization of shape information in a 3-D model description. Each box corresponds to a 3-D model; with its model axis on the left side of the box and the arrangement of its component axes are shown on the right side. In addition some component axes have 3-D models associated with them and this is indicated by the way the boxes overlap. The relative arrangement of each model's component axes, however, is shown improperly since it should be in an object-centred system rather than the viewer-centred projection used here (a more correct 3-D model is shown in figure 5). The important characteristics of this type of organization are: (i) each 3-D model is a self-contained unit of shape information and has a limited complexity, (ii) information appears in shape contexts appropriate for recognition (the disposition of a finger is most stable when specified relative to the hand that contains it), and (iii) the representation can be manipulated flexibly. This approach limits the representation's scope however, since it will only be useful for shapes that have well defined 3-D model decompositions.

very elaborate, only one 3-D model has to be dealt with at a time, and individual 3-D models have a limited complexity.

The coordinate system of the 3-D model

There are two kinds of object-centred coordinate system that the 3-D model representation might use. In one, all the component axes of a description, from torso to eyelash, are specified in a common frame based on the axis of the whole shape. The other uses a distributed coordinate system, in which each 3-D model has its own local coordinate system. We choose the latter for the following reasons: The spatial relations specified in a 3-D model description are always local to one of its models and should be given in a frame of reference determined by that model for the same reasons we prefer an object-centred system over a viewer-centred one. To do otherwise would cause information about the relative dispositions of a model's components to depend on the orientation of the model axis relative to the whole shape. For example, the description of the shape of a horse's leg would depend on the angle the leg makes with the torso. In addition to this stability and uniqueness consideration, the representation's accessibility and modularity is improved if each 3-D model maintains its own coordinate system, because it can then be dealt with as a completely self-contained unit of shape description.

The local coordinate system, for specifying the relative arrangement of a 3-D model's component axes, can be defined by its model axis or by one of its component axes. We shall refer to the axis chosen for this purpose as the model's *principal axis*. For the examples of this paper, the principal axis will be the component axis that meets or comes close to the largest number of other component axes in the 3-D model (for example the torso of an animal shape). The location of the principal axis must also be specified relative to the model axis, in order to maintain the connectedness of the distributed coordinate system.

Two three-dimensional vectors are required to specify the position in space of one axis relative to another. This can be done either by specifying the locations of its two end points or by specifying one end point and using the other to give its orientation. The second method will be more useful to us here. When one axis, S, is specified relative to another, A, it will be convenient to represent the location vector in cylindrical coordinates, (p, r, θ) where p is the position along the length of A (0 and 1 correspond to the endpoints of A), r is the radial distance away from A, and θ is the angular rotation about A as shown in figure 4a. The orientation vector will be specified in spherical coordinates, (i, ϕ, s), where i is the angle of inclination away from the direction of A; ϕ, like θ, is the rotation about A; and s is the size of S relative to A, as in figure 4b. We shall call the combined specification $(p, r, \theta, i, \phi, s)$ an *adjunct relation* for S relative to A.

Because of the varying precision with which 3-D models can represent a shape, it is appropriate to represent the angles and lengths that occur in an adjunct relation in a system that specifies both a value and a tolerance. For example, it is possible to state that a particular axis like the arm component of the human 3-D model in figure 3 is connected rather precisely at one end of the torso with ϕ coarsely specified and very little restriction on i. An example of such a system is illustrated in figure 5.

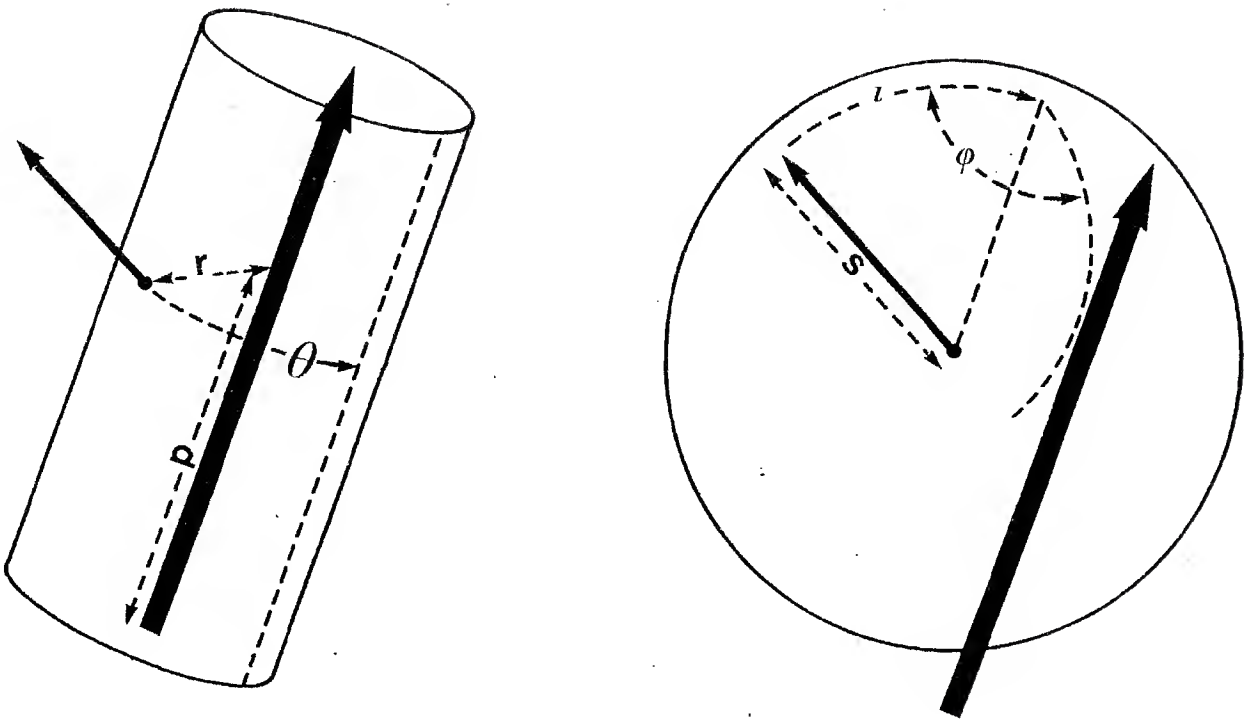
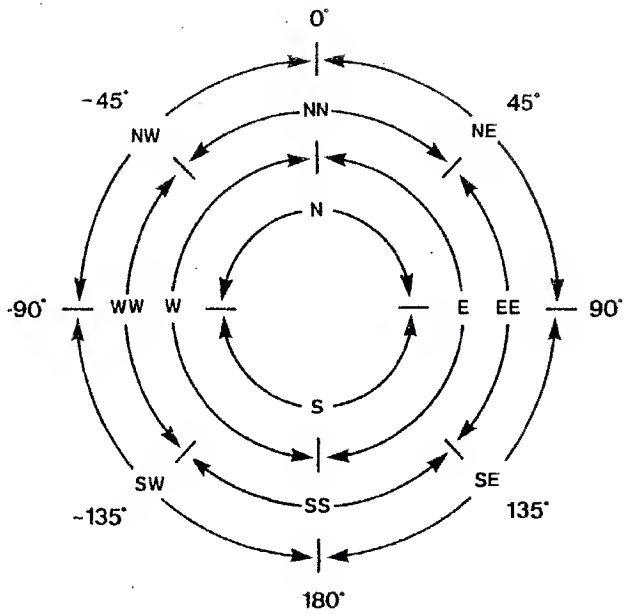
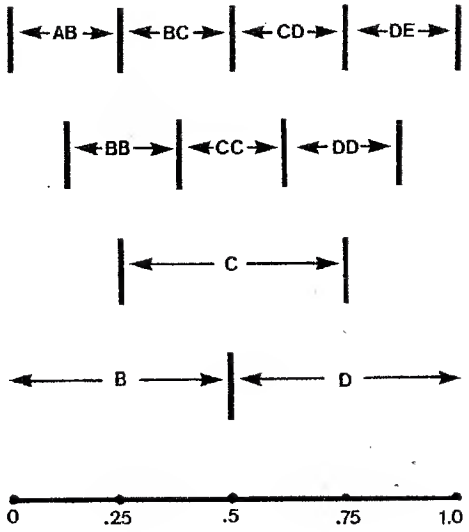


Figure 4. The spatial organization of a 3-D model's axes is specified in terms of pairwise relationships between those axes which we call *adjunct relations*. The disposition in space of one axis, S, is determined relative to another, A, by specifying the location of one of its end points in a cylindrical coordinate system, (p, r, θ), about A as shown on the left, and its orientation in a spherical coordinate system, (t, ϕ, s), centred on that point and aligned with A as shown on the right.

Figure 5. Angle and distance specifications in an adjunct relation must include tolerances so that the specificity of these parameters can be made explicit in the representation. One way to do this is shown in the upper diagrams which associate symbols with angular and linear ranges respectively. An example of adjunct relations for the human 3-D model in figure 3 using these symbols is shown in the lower table. *A* and *S* identify the two axes related by the adjunct relation specified on each row. If these mnemonic names were replaced by internal references to the corresponding 3-D models whenever they exist and left blank otherwise, this table would show essentially all the information carried by a 3-D model.





A	S	p	r	θ	i	φ	s
model	torso	BC	AB	NN	NN	NN	CC
torso	head	DE	AB	NN	NN	NN	BB
torso	arm	DE	BB	EE	E	E	DD
torso	arm	DE	BB	WW	E	W	DD
torso	leg	AB	BB	EE	SS	NN	DE
torso	leg	AB	BB	WW	SS	NN	DE

Deriving and using the 3-D model representation

The advantages of modularity, which have been one of our major concerns in the design of the 3-D model representation, will become especially visible as we discuss the processes that derive and use the representation for recognition. In particular, none of the processes has to deal with the internal details of more than one 3-D model at a time even if the complete description of a shape involves many 3-D models. We begin by examining the basic problems associated with identifying a model's coordinate system, its component axes, and transforming the viewer-centred axis specifications into specifications in the model's coordinate system. We then treat the task of recognizing this description as a problem of indexing a catalogue of stored 3-D model descriptions. Finally, we consider the interaction between the process that derives a 3-D model description and the recognition process. The ambiguities introduced by the perspective projection often mean that only coarse specifications of the lengths and orientations of a shape's axes are directly accessible from its image. However, if the recognition process proceeds conservatively in conjunction with the derivation process, it will be possible for it to make additional constraints available so that a more precise description can be produced.

Deriving a 3-D model description from an image

The construction of a 3-D model requires (i) the identification of the model's coordinate system and its component axes from an image and (ii) the specification of the arrangement of the component axis in that coordinate system.

Coordinate system and component identification

Even if a shape has a canonical coordinate system and a natural decomposition into component axes, there is still the problem of deriving them from an image. At present we do not have a complete solution to this problem, but some results have been obtained for shapes that fall within the scope of the 3-D model representation. Marr (1977) has shown that the image of a generalized cone's axis may be found from the occluding contours in an image, provided that the axis is not too foreshortened. An example of the decomposition formed by this method appears in figure 6, and a brief description is given in the legend. Notice that the final decomposition (f) was derived from the contour (a) without *a priori* knowledge of the three-dimensional shape apart from the assumption that it is composed of generalized cones. The method can therefore be used to find the component axes for the 3-D model of a shape that has not been seen before.

This result is somewhat limited, but so is the information it uses. The contours Marr studied are formed by rays that are tangential to the side of a smooth surface. (Interestingly, these particular contours are unsuitable for use in either stereopsis or structure-from-motion computations, because they do not correspond to fixed locations on the viewed surface.) Creases and folds on a surface also give rise to contours in an image, and these have yet to be studied in detail. Similarly much work remains in the study of how to use information about shape from shading, texture, stereo and motion.

A major difficulty in the analysis of images arises when an important axis is obscured, either because it is foreshortened or is hidden behind another part of the shape.

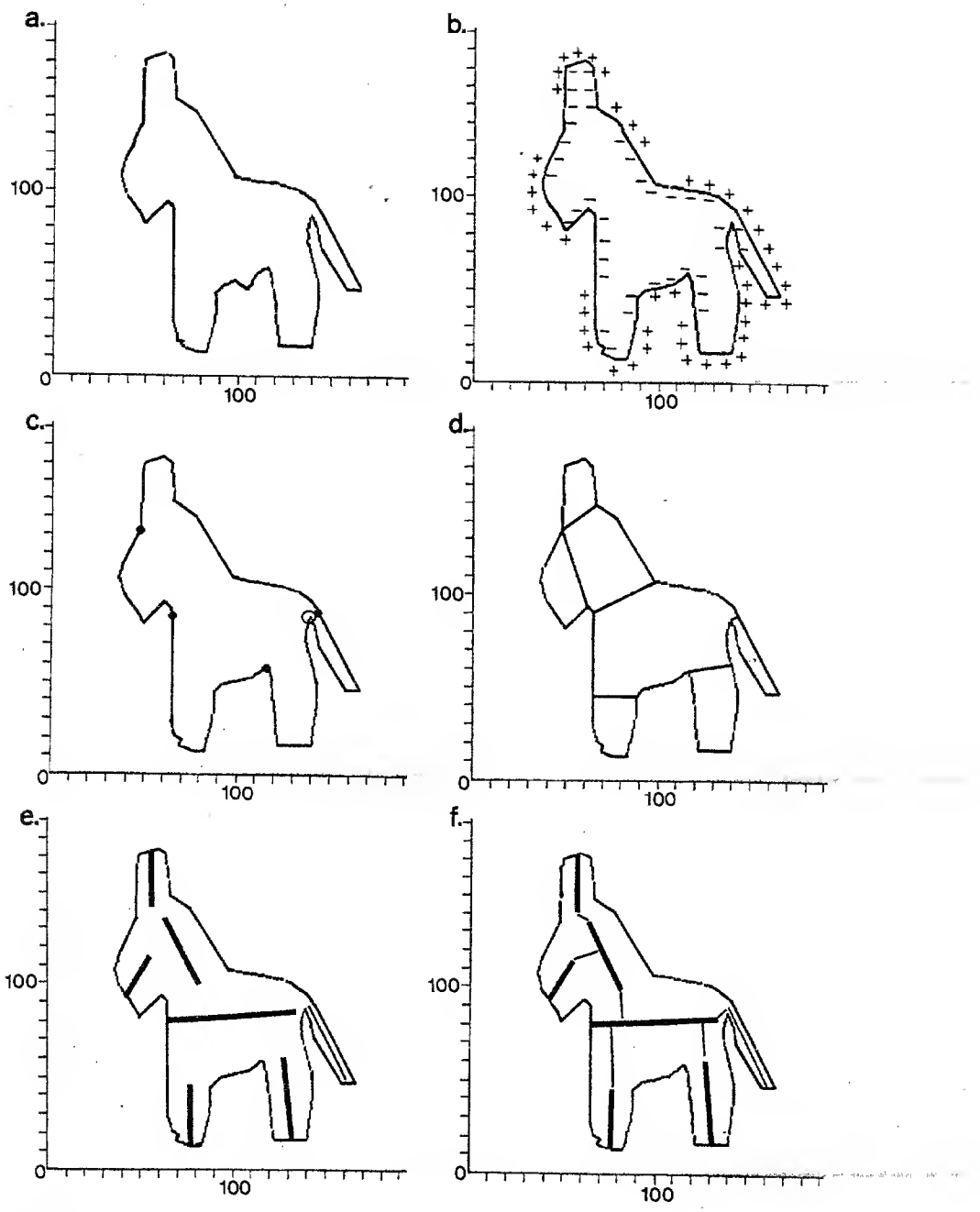


Figure 6. The occluding contours of simple shapes composed of generalized cones can be used to locate projections of its natural axes provided that they are not severly foreshortened (Marr 1977). One approach to doing this is shown in this example from a program written by P. Vatan. The initial outline in (a) was obtained by applying local grouping processes to the *primal sketch* of an image of a toy donkey (Marr 1976). This outline was then smoothed and divided into convex and concave sections to get (b). Next, strong segmentation points, like the deep concavity circled in (c), are identified and a set of heuristic rules are used to connect them with other points on the contour to get the segmentation shown in (d). The component axes shown in (e) were then derived from these. The thin lines in (f) indicate the position of the head, leg, and tail components along the torso axis, and the snout and ear components along the head axis.

For example, although the torso-based coordinate system for the overall shape of a horse is easily obtained from a side view, it is difficult to obtain when the horse faces the viewer. There are three ways of dealing with this situation. The first is to allow partial descriptions, that are based on the axes that can be seen from the front, to be used for recognition. If this were done, the representation would be slightly weakened in terms of the uniqueness criterion but not as severely as a purely viewer-centred representation. Another strategy is to use a shape's visible components whenever their recognition is easy but that of the overall shape is difficult. For example, the front view of a horse usually contains an excellent view of the horse's face which can be recognized directly and would provide another route by which the horse can be recognized. This strategy will be discussed further at the end of this section. Finally, it is sometimes possible to discover an axis that points directly at the viewer using radial symmetry in the image.

A water-bucket like that shown in figure 7 provides an interesting example of this. Its principal axis and the shape about that axis are derivable by the methods discussed above for the view shown in figure 7a but not for the view in figure 7c where the bucket's principal axis is foreshortened. An erroneous axis is likely to be established instead, perhaps going through the flanges that attach the handle to the rim. However, a failure to produce a recognizable description using this axis, would suggest that the correct axis is not the most pronounced in the image and an alternative can be sought. The two concentric circles (made by the top and bottom rims of the bucket) are strong clues which suggest that the principal axis passes through their centres. Furthermore, because they are concentric, these circles may be at widely separated locations along that axis and considering that possibility leads to the desired description of the bucket though the identity of the closer rim remains ambiguous. A local surface depth map like that illustrated in figure 2, computed using stereopsis, shading, or texture gradients for example, is likely to play an important role in interpreting images like these.

Relating viewer-centred to object-centred coordinates

Techniques for finding axes in a two-dimensional image describe their locations in a viewer-centred coordinate system, and so a transformation is required to convert their specifications to an object-centred coordinate system. In the 3-D model representation all axis dispositions are specified by adjunct relations, as in figure 4, so a mechanism is required for computing an adjunct relation from the specification of two axes in a viewer-centred coordinate system. We shall call this mechanism the *image-space processor*.

The image-space processor can be kept very simple. The adjunct relation is the only positional specification that has to be interpreted, since it is also how one links the coordinate systems of different 3-D models within the same description; furthermore, the number of adjunct relations within a 3-D model is small, and the image-space processor need deal with only one at a time. It will also be useful to compute the reverse transformation, from adjunct relation to two-dimensional projection, because this makes it possible to compute the appearance of a model's component axes given the orientation of the model's principal axis relative to the viewer. For either transformation, the image-space processor must maintain an object-centred coordinate frame specified in viewer-centred

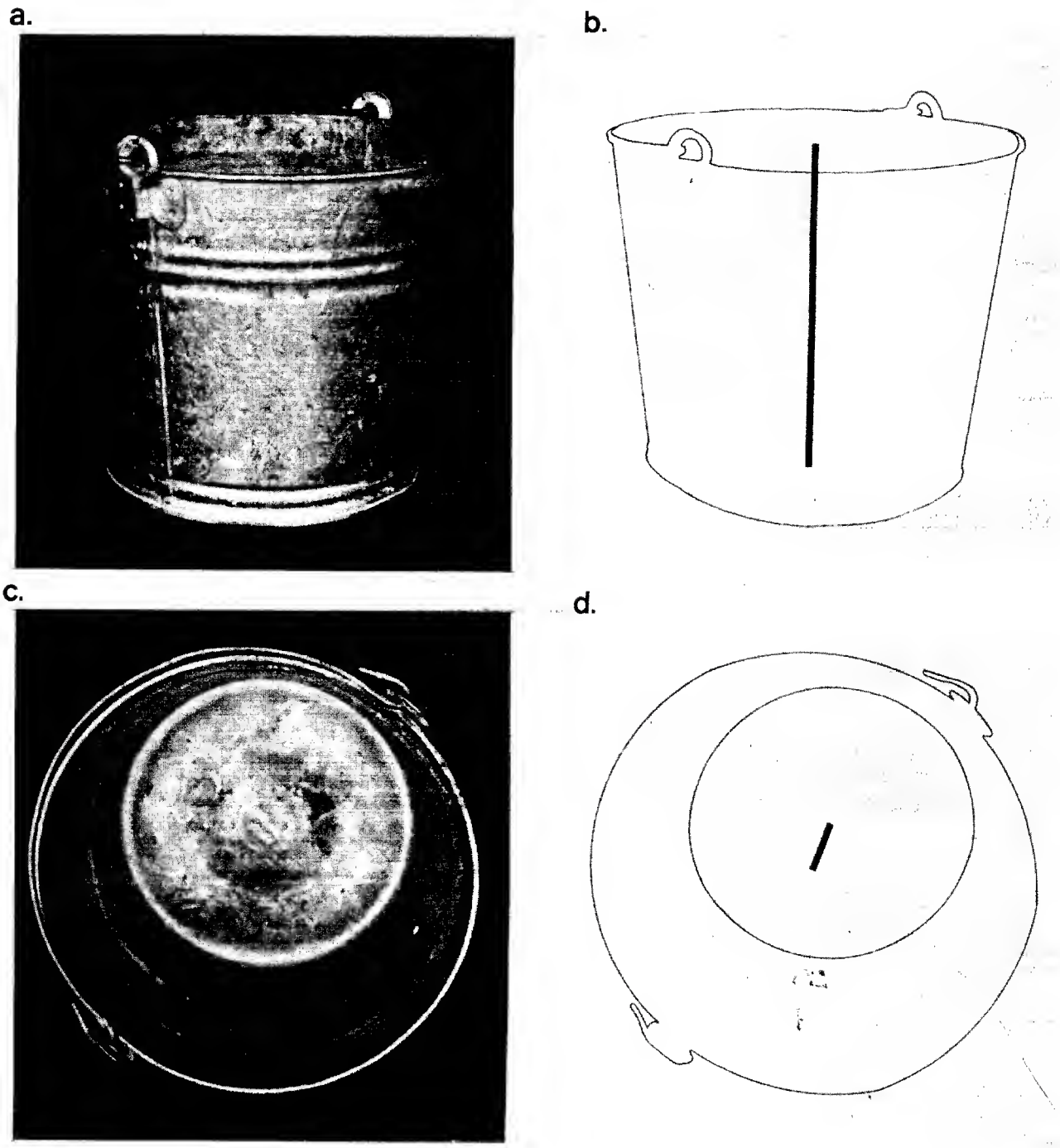


Figure 7. These views of a water bucket illustrate an important characteristic of any system based on the derivation of canonical axes from an image. The techniques useful for finding the axis shown in (b) from the image (a) are quite different from those that are best for situations where the axis is foreshortened as in (c) and (d).

coordinates. This frame is maintained by the image space processor in terms of an axis specification which we shall refer to as the *AXIS* and a vector which defines the direction of zero θ and ϕ about the AXIS. The AXIS is used to represent the position and orientation of a 3-D model's principal axis, and a second axis, which we call the *SPASAR*, is maintained to represent the disposition of a component axis of the model. The image-space processor makes the coordinates of the SPASAR available simultaneously in a frame centred on the viewer and in one centred on the AXIS, so that specifying the SPASAR in either frame makes it available in the other.

The accuracy of the adjunct relations computed by the image space processor is limited by the precision to which the AXIS and SPASAR are specified in the viewer-centred coordinate system. Since depth information is lost in the perspective projection, the precision of the orientation specifications for axes derived from the retinal images is least for the amount the axes dip towards or away from the viewer. Axis dip parameters can often be reconstructed at least roughly using stereopsis, shading, texture gradients, structure-from-motion, and contour analysis. Constraints supplied by the recognition process can also be used to improve the precision of the dip specifications. We shall consider this possibility later when we discuss the interaction between the derivation process and recognition.

Indexing and the catalogue of 3-D Models

Recognition involves two things, (i) a collection of stored 3-D model descriptions and (ii) various indexes into the collection that allow a newly derived description to be associated with a description in the collection. We shall refer to the above collection along with its indexing as the *catalogue of 3-D models*. Although our knowledge of what information can be extracted from an image is still limited, there are three access paths into the catalogue that appear to be particularly useful. They are the specificity index, the adjunct index, and the parent index.

3-D models can be classified hierarchically according to the precision of the information they carry, and an index can be based on this classification. We call it the *specificity index*, and figure 8 shows an example of this organization for models of a few animal shapes. The topmost level contains the most undifferentiated description available, a 3-D model without a component decomposition; only the model axis is specified so the model describes any shape. At the next level of detail, there is a general quadruped shape, a biped shape, a bird-like shape, and various limbs. These descriptions are most sensitive to the number of component axes in the model and to their distribution along the principal axis (which is the torso for most animal shapes), while only very coarse information about the lengths and orientations of the components is available. One level lower in the hierarchy the descriptions become more sensitive to angles and lengths, so that distinctions can be made for example between horse, giraffe, and cow shapes. A newly derived 3-D model may be related to a model in the catalogue by starting at the top of the hierarchy, and working down the levels through models whose shape specifications are consistent with the new model's, until a level of specificity is reached that corresponds to the precision of the information in the new model.

Once a 3-D model for a shape has been selected from the catalogue, its adjunct relations provide access to 3-D models for its components based on their locations, orientations and relative sizes. This gives us another access path to the models in the catalogue, and we call it the *adjunct index*. It tells one, for example, that the two similar components lying at the front end of a quadruped model are general limb models, and for a horse model, they are more specific horse-limb models. The adjunct index is useful for providing default information about the components of a shape prior to the derivation of 3-D models for them from the image, and also in situations where a catalogued model is not accessible *via* the specificity index because the description derived from the image is inadequate (perhaps because the component has very little structure).

The third access-path that we consider important is the inverse of the second and we call it the *parent index* of a 3-D model. When a component of a shape is recognized, it can provide information about what the whole shape is likely to be. For example, the catalogue's 3-D model for a horse can be indexed under each of its component 3-D models so that the 3-D model for a horse's leg provides access to the 3-D model for a horse-shape. This index would play an important role in the situation we discussed earlier, where an important axis of a shape is obscured or foreshortened. When a horse faces the viewer, the omission of the torso and hindleg axes might cause the neck axis to be selected incorrectly as the principal axis. Unless special provision has been made to handle this case, the specificity index will fail to access a horse model in the catalogue. A reasonable strategy to follow at this point would be to apply the derivation process to the components of the image. In this example, 3-D models for the head, neck, and the two forelegs would be produced. Catalogued models for the head and legs are likely to be found using the specificity index and each of these would indicate *via* the parent index that it is a component of either the quadruped or the horse 3-D model (depending on the quality of the derived component models), providing strong evidence for considering the quadruped or horse model for the whole shape.

It is important to note that the adjunct and parent indexes play a role secondary to that of the specificity index, upon which our notion of recognition rests. We shall see below that their purpose is primarily to provide contextual constraints that support the derivation process, for example by indicating where the principal axis is likely to be when such information cannot be obtained directly from the image. They do not prevent novel composite shapes such as a centaur from being described faithfully and from being recognized, for example, as a horse shape with a human bust.

It may be useful to construct other indexes into the catalogue, based for example on color or texture characteristics (e.g. the stripes of a zebra), or even on non-visual clues such as the sounds an animal makes, but these lie outside of the scope of this paper.

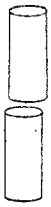
The interaction between derivation and recognition

So far, the derivation of a 3-D model has been treated separately from the process of relating that model to the stored models of the 3-D model catalogue. We view recognition as a gradual process proceeding from the general to the specific, that overlaps with, guides, and constrains the derivation of a description from the image. When a

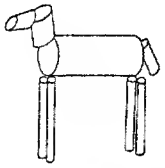
Figure 8. If the recognition process of relating new shape descriptions to known shapes is to be useful as a source of reliable information about the shape, it must be conservative. This diagram illustrates an organization (or indexing) of stored shape descriptions according to their specificity. The top row contains the most general shape description which carries information about size and overall orientation only. Since no commitment about the shape's internal structure is made, all shapes are described equally well. Descriptions in the second row include information about the number and distribution of component axes along the principal axis, making it possible to distinguish a number of shape configurations (a few are shown in this example). At this point only very general commitments are made concerning the relative sizes of the components and the angles between them. These parameters are made more precise at the third level so that distinctions can be made, for example, between the horse and cow shapes. A newly derived 3-D model would be related to a model in this catalogue by starting at the top level and working downwards as far as the information in the new description allows.

CYLINDER

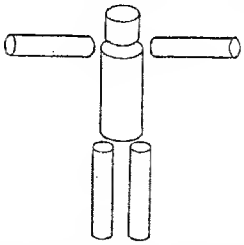
LIMB



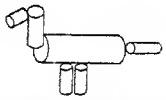
QUADRUPED



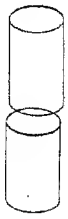
BIPED



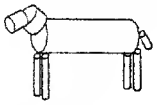
BIRD



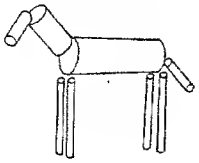
THICK LIMB



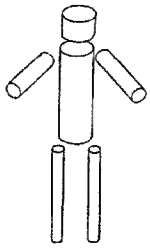
COW



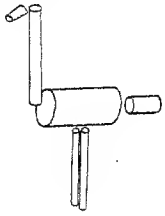
HORSE



HUMAN

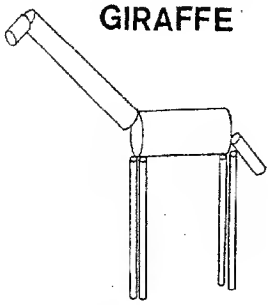


OSTRICH

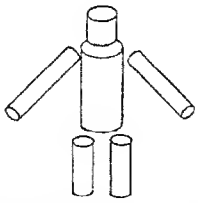


THIN LIMB

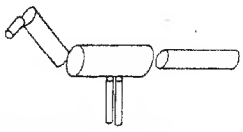
GIRAFFE



PRIMATE



DOVE



catalogued model is selected using one of the three indexes above, we want to use it to improve the analysis of the image. There are two phases to this, (i) the component axes from the image must be paired with the adjunct relations supplied by the catalogue, and (ii) the image space processor must be employed to combine the constraints available from the image with those provided by the model to produce a new set of derived adjunct relations that are more specific than those from the catalogue model. This last phase involves an analysis of constraints that derives adjunct relations consistent with both the image and the information from the catalogue. The general idea of using a stored model of a shape to assist in the interpretation of an image was first used by Roberts (1965) in a computer program for describing images of shapes built out of cubes, wedges, and hexagonal prisms in terms of their edges.

Finding the correspondence between image and catalogued model

The first phase can be thought of as a homology problem, in which the adjunct relations of a catalogue model must be related to the axes derived from an image. There may not be a complete solution. For example, the leg axes in a silhouette of a horse from the side are easily identified, but the left and right forelegs cannot usually be disambiguated without further information. Often this may be tolerable however, since the corresponding adjunct relations for the two legs have the same general orientation specifications (they differ only in their locations), and this is all that the following analysis makes use of.

The information available for establishing the correspondence between image and model increases as the derivation-recognition process proceeds. Initially, positional information along the principal axis of the stick figure has priority since it is the least distorted by the perspective projection. Other clues available initially include (i) the relative thicknesses of the shapes about the component axes (the neck of a horse is much thicker than the legs); (ii) possible decompositions of component axes (the tail and legs of a horse may be roughly straight, but the bust has two components that always make a large angle with one another); (iii) symmetry or repetition (the legs of a horse are all the same thickness, are roughly parallel, and because of this have roughly the same length and orientation in the image, distinguishing them from the tail); and (iv) large differences in the ϕ angle of the adjunct relation (in an image, the legs and tail of a horse usually extend to one side of the torso while the bust extends to the other). Collectively, such clues are often sufficient to relate the major components of a 3-D model to the axes derived from an image.

Homology information is also made available by the adjunct and parent indexes. When a 3-D model from the catalogue is obtained using the adjunct index the polarity of that components axis is automatically determined. For example, when continuing the analysis of the image of a horse to one of the legs, the polarity of the leg axis is indicated by its connection with the torso (the hoof end being distal to that junction). When the selection of a catalogue model is based on the identification of a shape's components using the parent index, the pairings for these components are already determined and they strongly constrain pairings for the remaining components. For example, in the case of the

horse facing the viewer, the missing torso's location in the image can be found from the locations of the head, neck and forelegs.

Constraint analysis

Once a homology has been established between a 3-D model and the image, we want to use the additional information it makes available to constrain the possible dip angles for the axes. The basic idea is that there are often only a few combinations of dip specifications for the projected axes in the image for which the adjunct relations derived from the image would be consistent with those supplied by the catalogue model. Or equivalently, there are often only a few orientations of the catalogue model's principal axis (relative to the viewer) for which the appearance of its component axes match closely the projected axes in the image.

The combination of information from the image and the catalogue model is often sufficient to determine the axis dips uniquely up to reflexion. For example in figure 9, (a) shows the locus of AXIS orientations (relative to the viewer) that are consistent with an inclination of 90 degrees between the AXIS and the SPASAR, and an angle of 47 degrees between their projections onto the image plane; (b) shows the allowed orientations for an inclination angle of 45 degrees and a projected angle of -111 degrees; and (c) shows the intersection of these two sets. The sharpness of these constraints depends on the particular viewing angle (as indicated by the other examples in figure 9), and on the particular adjunct relations in the model. Generally, the constraints are the strongest when the component axes have very different orientations and when the principal axis does not lie in the image plane.

There are several algorithms that can use these constraints. Perhaps the simplest is a *relaxation* process that adjusts the orientation of the AXIS incrementally, seeking the disposition for which the projections of the angles between the component axes of the catalogue model, as computed by the image space processor, best agree with those in the stick figure image. At this point the AXIS indicates the orientation of the principal axis that is most consistent with all of the constraints and the SPASAR can be used to compute the orientations of each of the component axes using the adjuncts from the catalogue model. This hill climbing approach converges quite efficiently when the constraints are sufficiently strong. Alternatively, instead of relaxing the orientation of the catalogue model's principal axis, one can relax the dip angles of the sticks obtained from the image. In this case, the discrepancy measure is obtained by comparing adjunct relations derived between the sticks in the image with the corresponding adjunct relations from the catalogue. This approach is interesting because in its implementation, all of the transformations carried out by the image space processor are in the same direction (from viewer-centred to object-centred coordinates). In a final step, improved orientation information may be used to recover more information from the image. In particular, once the orientations of the axes have been determined, their relative lengths may be computed.

The overall recognition process may be summarized as follows. One first selects a model from the catalogue based on the distribution of components along the length of the principal axis. This model then provides relative orientation constraints that help to

Figure 9. Specifying the inclination angle, i, that the SPASAR makes with the AXIS, and the angle between their images strongly constrains the orientation of the AXIS' coordinate system relative to the viewer. (a) shows the orientations consistent with an inclination of 90 degrees and an image angle like that between the heavy lines in the accompanying stick figure (allowing a tolerance of plus or minus five degrees in the image angle). The horizontal axis of the graph indicates the angle the AXIS dips out of the image plane towards the viewer. The vertical axis is the amount the coordinate system is rotated about the AXIS. (b) shows the set of orientations consistent with i = 45 degrees and the angle between the images of the torso and neck axes. (c) shows the intersection of the two sets which is restricted to a narrow range of orientations having a dip of approximately 67 degrees out of the image plane (there is another solution not shown here at -67 degrees). The remaining rows show cases for a dip of 45 degrees and zero degrees respectively.

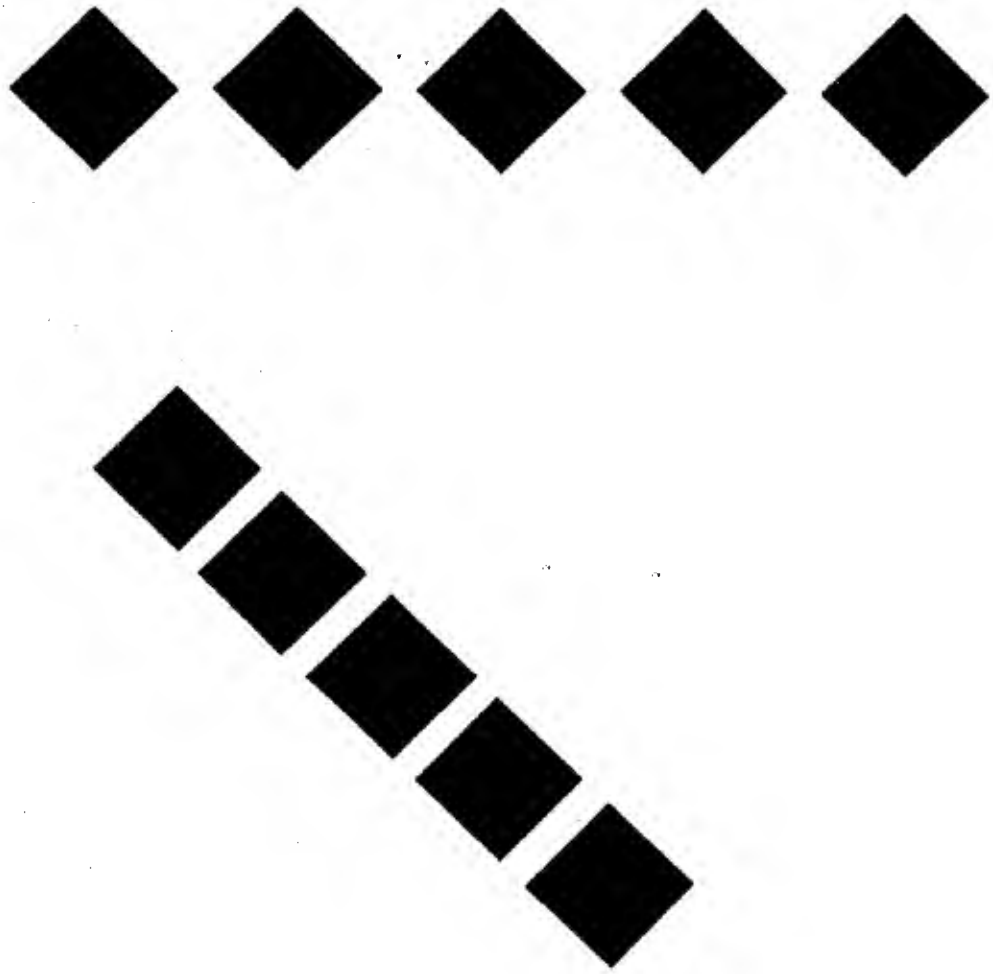


Figure 10. The effect of different choices of an object-centred coordinate system on the perception of shape is apparent in these diagrams which were adapted from Attneave (1968). The black shapes can be seen as diamonds or squares depending on which of their several natural axes is used.

determine the absolute orientations (relative to the viewer) of the component axes in the image, and with this information the image space processor can be used to compute their relative lengths. This new information can then be used to disambiguate shapes at the next level of the specificity index.

Discussion

We have outlined a theory for visual shape recognition that is based on the specification of a representation for shape and a discussion of the nature of the processes that derive and use it. It is incomplete both in terms of its limited scope of applicability and in the specification of many of its details; but it raises several precise questions that are apparently important for shape recognition. In this section we consider the theory in a broader perspective, examining its relationship to other parts of the visual process, and some of the questions it raises for psychology.

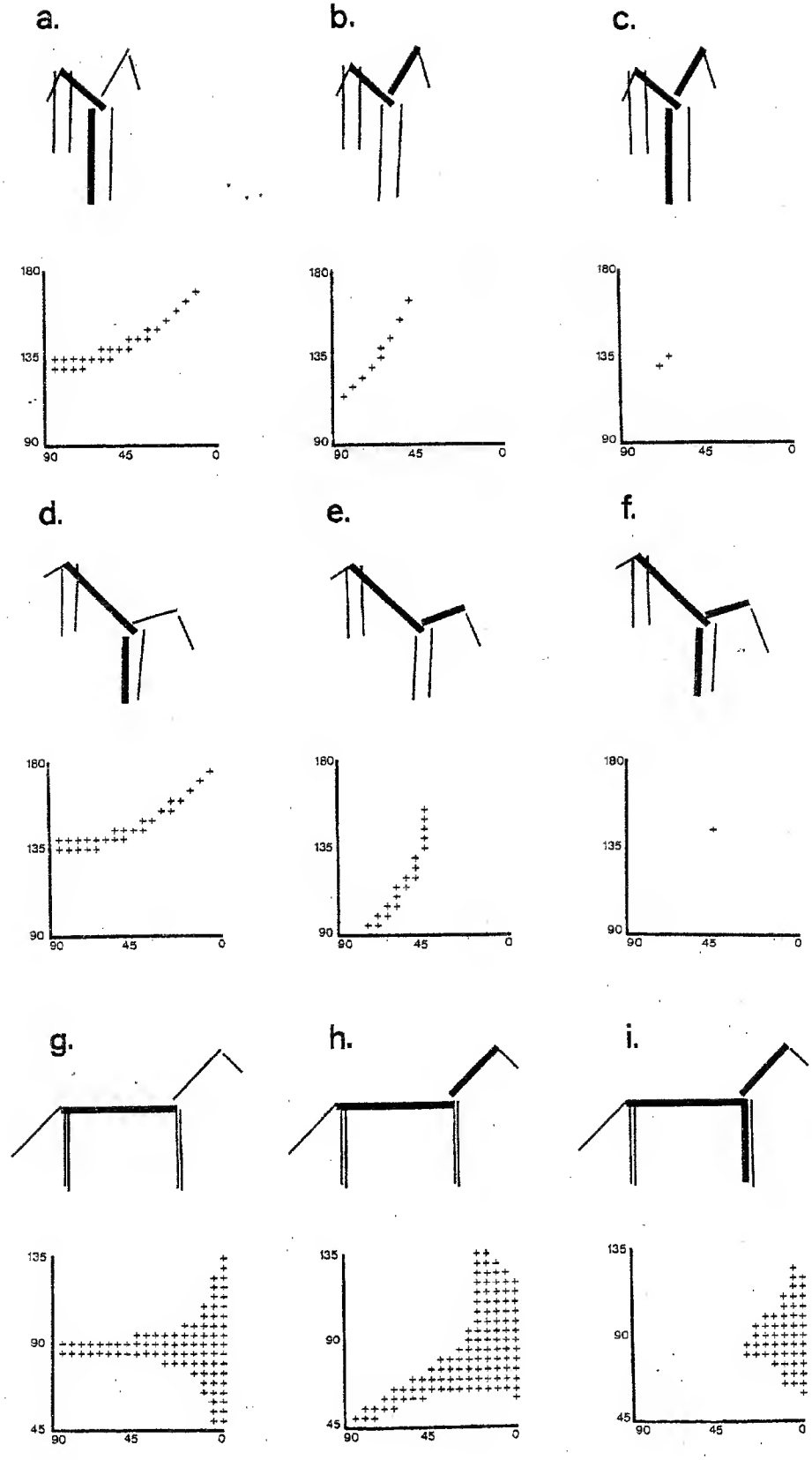
The approach

We have studied vision as a process that assembles descriptions in a number of representations, each specializing in some aspect of the visual scene with later ones building on the information made explicit by those before them. This approach is suggested by several experimental findings (for example, Hubel & Wiesel 1962), and it is consistent with the *principal of modularity* (Marr 1976) which states that any large computation should be split up into a collection of small, nearly independent, specialized sub-processes. If visual information processing were not organized in this way, incremental changes in its design would be unable to improve one aspect of the process' performance without simultaneously degrading the operation of many others.

There are several criteria that help us to identify modules in the visual process. Early in the processing, computational considerations (what can be computed from the available information) and evidence from neurophysiology and psychophysics provide the most useful constraints. An example of an early representation is the *primal sketch* (Marr 1976), which represents the intensity changes and the local geometry of an image. For later modules like the one addressed in this paper, the strong constraints arise from what the representation is to be used for.

Psychological considerations

The ideas we have discussed can be examined experimentally at two levels, by considering either the type of information made explicit by the visual process in its representations, or by considering the nature of the processes that derive and maintain it. The first is the more fundamental; is a three-dimensional representation used, does it have a modular organization, and is it object-centred? These questions have yet to be put to empirical test, but three observations are worth noting here. The first is that stick figure animals like those shown in figure 1 are usually recognized easily despite the limited amount of shape information they portray. While this does not demonstrate that the human visual process is based on stick figures, it does suggest that the type of information carried by stick figures plays an important role in it.


a.
b.
c.
180
135
90
90
45
0
d.
e.
f.
g.
h.
i.
45

Secondly, illusions like that shown in figure 10 (due originally to Ernst Mach, and adapted from Attneave 1968) provide evidence that local shape information is described relative to axes that are defined more globally. In the top row, the shapes are seen as diamonds, whereas along the diagonal they are seen as squares. The diagonal axis is therefore being constructed during the analysis of this pattern; it influences, and therefore probably precedes, the description of the shapes of the local elements.

Thirdly, Warrington and Taylor (1973) drew attention to the difficulty experienced by their patients with right parietal lesions at interpreting certain views of common objects which they called *unconventional views*. For example, these patients would fail to recognize the top view of a bucket (figure 7c), denying that it was a bucket even when told that it was. The patients were relatively unimpaired on views like figure 7a. As Warrington and Taylor pointed out, one cannot easily explain this difference in terms of familiarity or impaired depth perception because both views of a bucket are common and depth is just as important to the three dimensional structure of figure 7a as it is of 7c. However, if the internal shape representation used for recognition were based on a shape's natural axes, the second figure would be more difficult to describe correctly since its major axis is foreshortened. If this explanation were correct, Warrington and Taylor's unconventional views would correspond to views in which an important natural axis of the shape is foreshortened in the image, making it difficult for the patient to discover or derive a description in the shape's canonical coordinate system.

Acknowledgements: We thank Whitman Richards for valuable criticism and especially our referee who helped us to improve the form of this presentation substantially. We are also indebted to Drew McDermott, Tomaso Poggio, and Kent Stevens who gave us comments on earlier versions of this paper, and to Karen Prendergast for preparing the drawings. This article describes work reported in M.I.T. A.I. Lab. Memos 341 and 377, and it was conducted at the Artificial Intelligence Laboratory, a Massachusetts Institute of Technology research program supported in part by the Advanced Research Projects Agency of the Department of Defense and monitored by the Office of Naval Research under Contract number N00014-75-C-0634.

References

Attneave, F. 1968 Triangles as ambiguous figures. *Amer. J. Psychol. 81*, 447-453.

Binford, T. O. 1971 Visual perception by computer. Presented to the IEEE Conference on Systems and Control, Miami, in December 1971.

Blum, H. 1973 Biological shape and visual science, (part I). *J. theor. Biol., 38*, 205-287.

Hubel, D. H. & Wiesel, T. N. 1962 Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol., Lond. 160*, 106-154.

Marr, D. 1976 Early processing of visual information. *Phil. Trans. Roy. Soc. B 275*, 483-524.

Marr, D. 1977. Analysis of occluding contour. *Proc. R. Soc. B* (in the press).

Marr, D. & Poggio, T. 1977 From understanding computation to understanding neural circuitry. In *The Visual Field: Psychophysics and Neurophysiology. Neurosciences Research Program Bulletin*, E. Poeppel *et. al.*, Eds. (in the press).

Minsky, M. 1975 A framework for representing knowledge. In: *The psychology of computer vision*, Ed. P. H. Winston, pp 211-277. New York: McGraw-Hill.

Roberts, L. G. 1965 Machine perception of three-dimensional solids. In *Optical and Electrooptical Information Processing*. J. T. Tippett *et. al.*, Eds. pp 159-197. Cambridge: M.I.T. Press.

Warrington, E. K. & Taylor, A. M. 1973 The contribution of the right parietal lobe to object recognition. *Cortex, 9*, 152-164.
Also remarks made by E. K. W. in a lecture given on Oct. 26th 1973 at the M.I.T. Psychology Department.

Winston, P. H. 1975 Learning structural descriptions from examples. In: *The psychology of computer vision*, Ed. P. H. Winston, pp 157-209. New York: McGraw-Hill.